Q1. Write about Positional Reference ?

Ans 1. Pig has a unique feature called positioning references. Assume a situation in which there are thousands of attributes then it's really tough to remember each attributes name. Pig allows the user to retrieve data according to its position. It is somewhat similar to array as we can retrieve data of array using A[0], A[1], it can retrieve data as $0, $1 position.

Q2. Write steps to count a field ?

Ans 2. There are vital three steps to count any field in pig.

1. First Load data to a bag.

grunt> chars = LOAD '/home/training/NewData2/data/b.txt' AS (letter:chararray);

dump chars;

2. Then group the field according to which you want to count.

grunt>groupLetters = group letters by letter;

dump groupLetters;

3. Then apply count command to the grouped data.

counts = foreach groupLetters generate group, COUNT(letters);

dump counts;

Q3. Explain Joins ?

Ans 3. There are various forms of joins. 'Joins' means adding two data sets using common attributes or keys.

1. Inner Join - It is also called default join. It gives intersection. It gives the common fields of two data sets.

SELECT <select\_list>

FROM Table AA

INNER JOIN Table BB

ON A.Key = B.Key

2. Outer Join - It gives union of two data sets.

SELECT <select\_list>

FROM Table AA

FULL OUTER JOIN Table BB

ON A.Key = B.Key

3. Left Join - It gives all the values of data set A if A is left data set & gives the values of B which is common to A.

SELECT <select\_list>

FROM Table AA

LEFT JOIN Table BB

ON A.Key = B.Key

4. Right Join - It gives all the values of data set B if B is right data set & gives the values of A which is common to A.

SELECT <select\_list>

FROM Table AA

FULL OUTER JOIN Table BB

ON A.Key = B.Key

Q4. Difference between local and Map Reduce Mode ?

Ans 4. Local mode: All scripts are run on a single machine without requiring Hadoop MapReduce and HDFS. This can be useful for developing and testing Pig logic. If you’re using a small set of data to developer or test your code, then local mode could be faster than going through the MapReduce infrastructure. Local mode doesn’t require Hadoop. When you run in Local mode, the Pig program runs in the context of a local Java Virtual Machine, and data access is via the local file system of a single machine. Local mode is actually a local simulation of MapReduce in Hadoop’s LocalJobRunner class.

Map Reduce mode (also known as Hadoop mode): Pig is executed on the Hadoop cluster. In this case, the Pig Script gets converted into a series of MapReduce jobs that are then run on the Hadoop cluster. If you have a terabyte of data that you want to perform operations on and you want to interactively develop a program, you may soon find things slowing down considerably, and you may start growing your storage.

Q5. Write 3 errors while practicing ?

Ans 5. 1. If you are typing commands of pig in localroot ( [@training]$ ), it will show error. You have to switch to pig & type in grunt [ grunt>].

Follow pig -x local to switch to pig.

2. While writing this flatten command, we cannot modify it to $1, $2. If done so, error is shown in form of log files.

flatBag = foreach tokenBag generate flatten($0);

3. While writing this substring command, if you give a hypothetical value which is far greater than the length of the string then it gives error. Ex- (token,0,15) gives error.

letters = foreach tokens GENERATE SUBSTRING(token,0,1) AS letter:chararray;